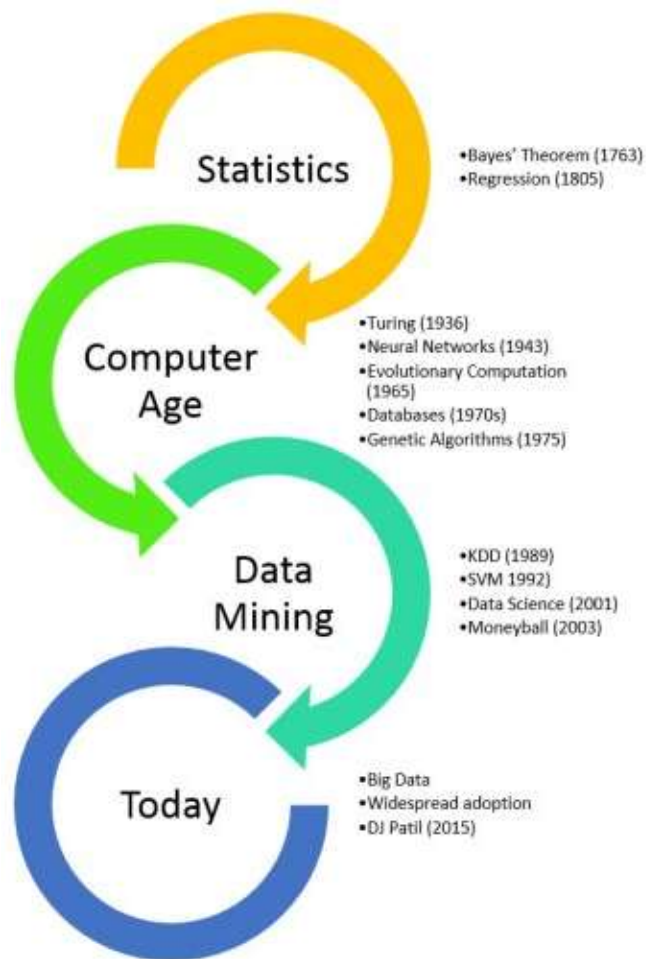


# Data Preprocessing in Data Mining

Dr. Tran Anh Tuan,  
Faculty of Mathematics and Computer Science,  
University of Science, HCMC

Dr. Tran Anh Tuan, Faculty of Mathematics and Computer  
Science, University of Science, HCMC



Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

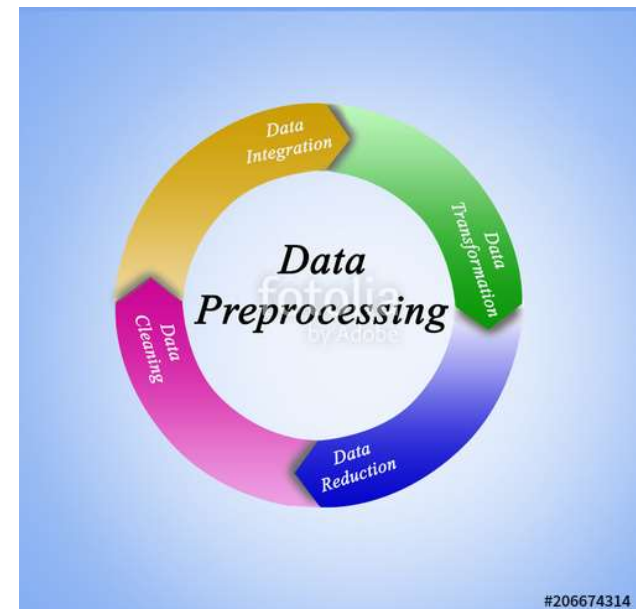
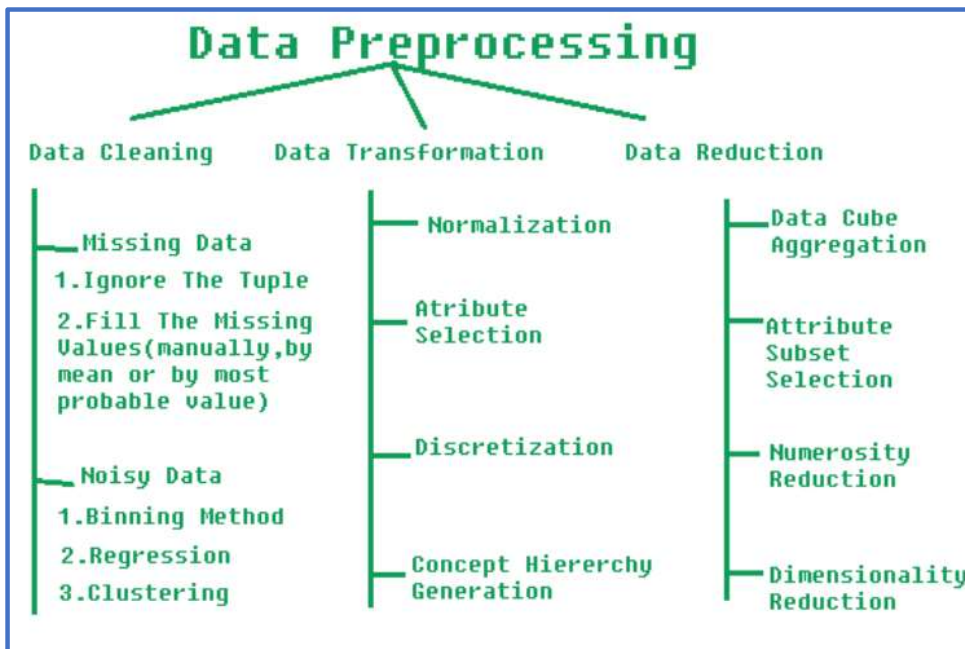
Dr. Tran Anh Tuan, Faculty of Mathematics and Computer Science, University of Science, HCMC

# Syllabus

- **Lecture 1 : Data Preprocessing**
- Lecture 2 : Explanatory Data Analysis
- Lecture 3 : Feature Engineering (Feature Importance and Selection)
- Lecture 4 : Association Rule Learning
- Lecture 5 : Unsupervised Clustering
- Lecture 6 : Unsupervised Clustering (cont.)
- Lecture 7 : Anomaly and Outlier Detection
- Lecture 8 : Regression and Classification Learning
- Lecture 9 : Recommendation Learning
- Lecture 10 : Final Project Requirement

# Introduction

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



# Data Cleaning

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

## (a). Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways.

Some of them are:

### 1. **Ignore the tuples:**

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

### 2. **Fill the Missing values:**

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.



# Missing Data

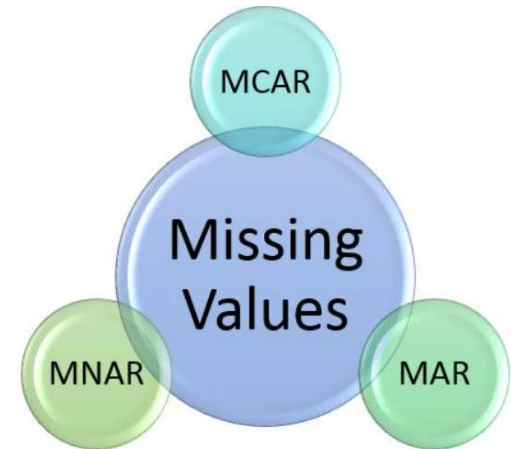
- What is missing data?
  - Missing data are defined as values that are not available and that would be meaningful if they are observed. Missing data can be anything from missing sequence, incomplete feature, files missing, information incomplete, data entry error etc.
  - Most datasets in the real world contain missing data. Before you can use data with missing data fields, you need to transform those fields so they can be used for analysis and modelling
  - Understanding the data and the domain from which it comes is very important.



# Missing Data

Missingness is broadly categorized in 3 categories:

- **Missing Completely at Random (MCAR)**
- **Missing at Random (MAR)**
- **Missing not at Random (MNAR)**



## **Missing Completely at Random (MCAR)**

When we say data are **missing completely at random**, we mean that the missingness has nothing to do with the observation being studied

For example, a weighing scale that ran out of batteries, a questionnaire might be lost in the post, or a blood sample might be damaged in the lab

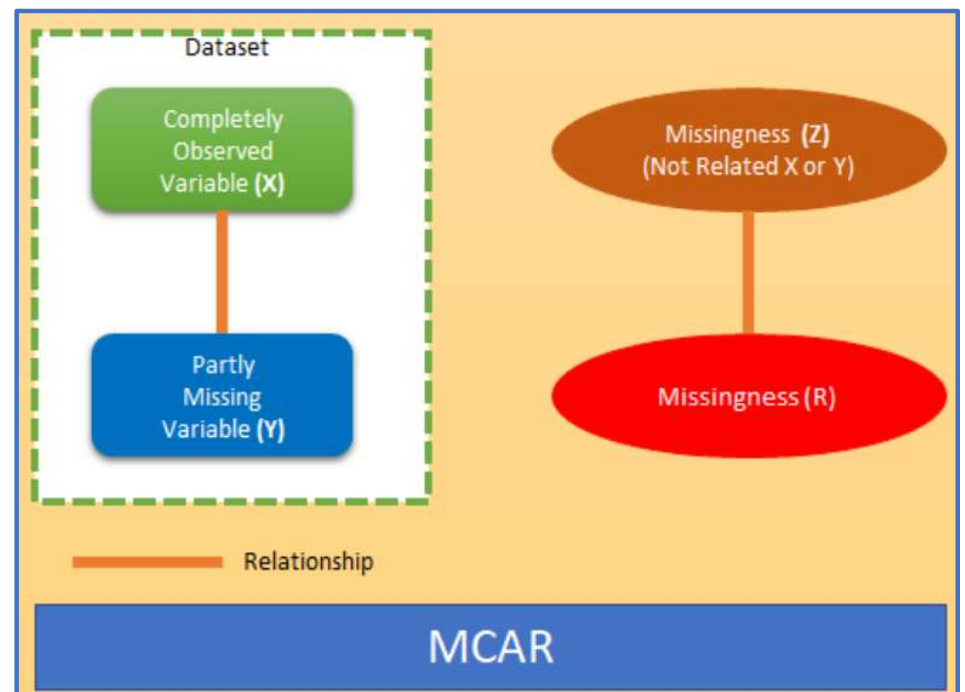
MCAR is an ideal but unreasonable assumption

# Missing Data

- MCAR Missing Data Example

<div><div>X</div><div>Y</div><div>R</div><div>Z</div></div>			
Mobile ID	Mobile Package	Download Speed	Missing because of data limit fully utilized
1	Fast+	157	
2	Lite	99	
3	Fast+	167	
4	Fast+	N/A	

It might be able to predict from other variable





# Missing Data

## Missing at Random (MAR)

- When we say data are **missing at random**, we mean that missing data on a partly missing variable ( $Y$ ) is related to some other completely observed variables( $X$ ) in the analysis model but not to the values of  $Y$  itself

It is not specifically related to the missing information.

**For example**, if a child does not attend an examination because the child is ill, this might be predictable from other data we have about the child's health, but it would not be related to what we would have examined had the child not been ill.

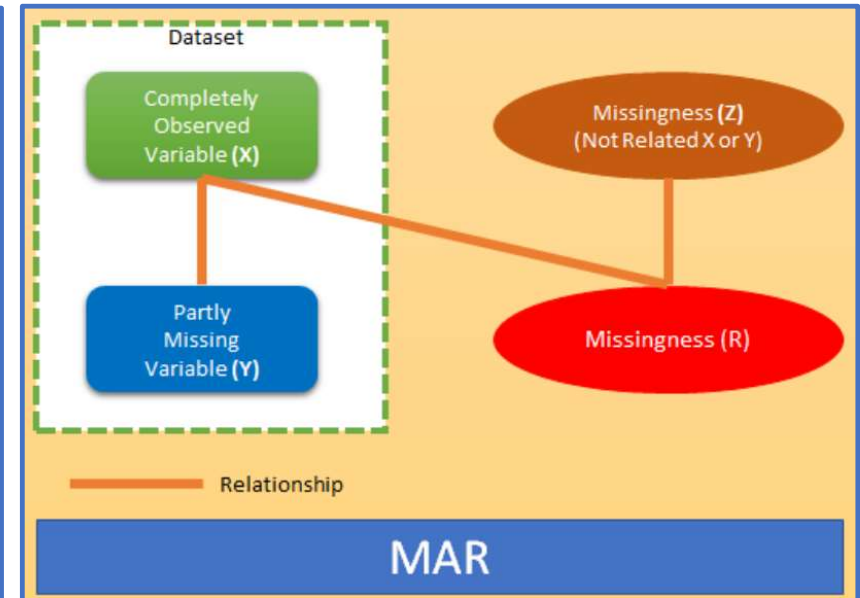
# Missing Data

- MAR Missing Data Example

				X	Y
Mobile ID	Mobile Package	Download Speed	Data Limit Usage		
1	Fast+	157	80%		
2	Lite	99	70%		
3	Fast+	167	10%		
4	Fast+	N/A	100%		

When Data limit Usage reached 100%, missing has occurred, Missing depends on other observed variable

It might be able to predict using observed variables



# Missing Data

## Missing not at Random (MNAR)

- When data are **missing not at random**, the missingness is specifically related to what is missing
- e.g. a person does not attend a drug test because the person took drugs the night before, a person did not take English proficiency test due to his poor English language skill.
- The cases of MNAR data are problematic.
- A pictorial view of MNAR as below where missingness has direct relation to variable Y. It can have other relation as well (X & Z)

The only way to obtain an unbiased estimate of the parameters in such a case is to model the missing data but that requires proper understanding and domain knowledge of the missing variable

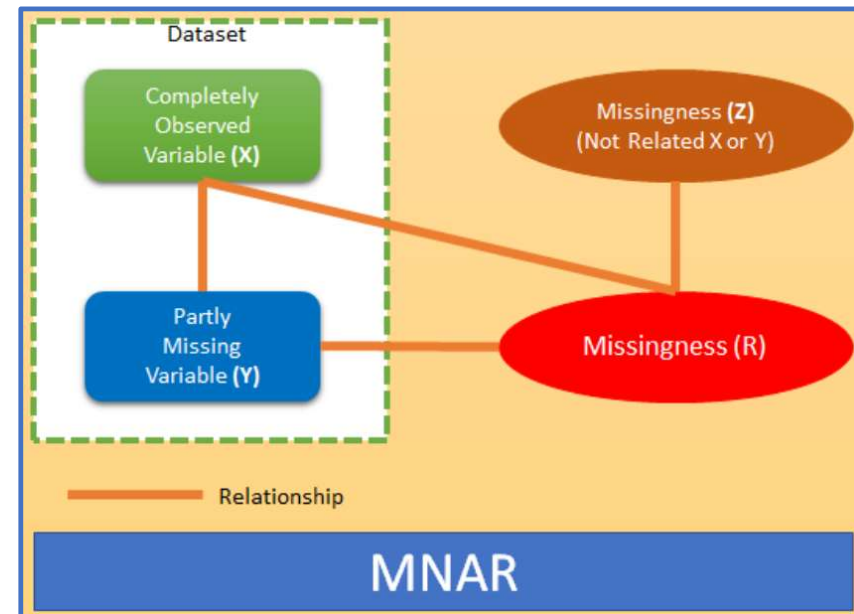
# Missing Data

- MNAR Missing Data Example

				X	Y
Mobile ID	Mobile Package	Download Speed	Data Limit Usage		
1	Fast+	N/A	80%		
2	Lite	99	70%		
3	Fast+	167	10%		
4	Fast+	N/A	75%		

Download speed missing value has relation to Download Speed, Data Limit Usage and some other unknown variable. Here value is missing beyond a data limit usage range ( $\geq 75\%$ ) but we can not predict the value

It is difficult to predict missing values



# Missing Data Solution

## List-wise (Complete-case analysis — CCA) deletion

- By far the most common approach to the missing data is to simply omit those cases with the missing data and analyse the remaining data
- If there is a large enough sample, where power is not an issue, and the assumption of MCAR is satisfied. The listwise deletion may be a reasonable strategy
- However, when there is not a large sample, or the assumption of MCAR is not satisfied, the listwise deletion is not the optimal strategy

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	N/A	N/A	95%
8	Lite	76	77%
9	Fast+	180	N/A

← Delete

← Delete

← Delete



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
5	Lite	76	70%
6	Fast+	155	10%
8	Lite	76	77%

# Missing Data Solution

## Pairwise (available case analysis — ACA) Deletion

- In this case, only the missing observations are ignored and analysis is done on variables present
- Pairwise deletion is known to be less biased for the MCAR or MAR data
- However, if there are many missing observations, the analysis will be deficient

Mobile ID	Mobile Package	Download Speed	Data Limit Usage	
1	Fast+	157	80%	
2	Lite	99	70%	
3	Fast+	167	10%	
4	Fast+	N/A	80%	Delete
5	Lite	76	70%	
6	Fast+	155	10%	
7	N/A	N/A	95%	Delete
8	Lite	76	77%	
9	Fast+	180	N/A	Delete

↓

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+		80%
5	Lite	76	70%
6	Fast+	155	10%
7			95%
8	Lite	76	77%
9	Fast+	180	



# Missing Data Solution

## Dropping Variables

- If there are too many data missing for a variable it may be an option to delete the variable or the column from the dataset
- There is no rule of thumbs for this but depends on situation and a proper analysis of data is needed before the variable is dropped all together
- This should be the last option and need to check if model performance improves after deletion of variable.

Delete



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	N/A	80%
2	Lite	N/A	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	N/A	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	77%



Mobile ID	Mobile Package	Data Limit Usage
1	Fast+	80%
2	Lite	70%
3	Fast+	10%
4	Fast+	80%
5	Lite	70%
6	Fast+	10%
7	Fast+	95%
8	Lite	77%
9	Fast+	77%


# Missing Data Solution

## Mean, Median and Mode

- In this imputation technique goal is to replace missing data with statistical estimates of the missing values. **Mean, Median or Mode** can be used as imputation value.
- In a mean substitution, the mean value of a variable is used in place of the missing data value for that same variable. This has the benefit of not changing the sample mean for that variable
- The theoretical background of the mean substitution is that the mean is a reasonable estimate for a randomly selected observation from a normal distribution
- However, with missing values that are not strictly random, especially in the presence of a great inequality in the number of missing values for the different variables, the mean substitution method may lead to inconsistent bias. Distortion of original variance and Distortion of co-variance with remaining variables within the dataset are two major drawbacks of this method

# Missing Data Solution

Mean (Download Speed) = 130




Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	130	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	130	95%
8	Lite	76	77%
9	Fast+	180	95%

Median (Download Speed) = 155



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	155	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	155	95%
8	Lite	76	77%
9	Fast+	180	95%

Median can be used when variable has a skewed distribution.

# Missing Data Solution

The rationale for Mode is to replace the population of missing values with the most frequent value, since this is the most likely occurrence.

Mode (Download Speed) = 200



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	200	80%
2	Lite	100	70%
3	Fast+	200	10%
4	Fast+	N/A	80%
5	Lite	50	70%
6	Fast+	200	10%
7	Fast+	N/A	95%
8	Lite	200	77%
9	Fast+	180	95%



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	200	80%
2	Lite	100	70%
3	Fast+	200	10%
4	Fast+	200	80%
5	Lite	50	70%
6	Fast+	200	10%
7	Fast+	200	95%
8	Lite	200	77%
9	Fast+	180	95%

# Missing Data Solution

## Last Observation Carried Forward (LOCF)

- If data is time-series data, one of the most widely used imputation methods is the last observation carried forward (LOCF). Whenever a value is missing, it is replaced with the last observed value. This method is advantageous as it is easy to understand and communicate.
- Although simple, this method strongly assumes that the value of the outcome remains unchanged by the missing data, which seems unlikely in many settings.

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	N/A	86%
6	6-Jan	155	87%
7	7-Jan	N/A	89%
8	8-Jan	N/A	90%
9	9-Jan	180	92%



Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	90	86%
6	6-Jan	155	87%
7	7-Jan	155	89%
8	8-Jan	155	90%
9	9-Jan	180	92%

# Missing Data Solution

## Next Observation Carried Backward (NOCB)

- A similar approach like LOCF which works in the opposite direction by taking the first observation after the missing value and carrying it backward (“next observation carried backward”, or NOCB).

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	N/A	86%
6	6-Jan	155	87%
7	7-Jan	N/A	89%
8	8-Jan	N/A	90%
9	9-Jan	180	92%



Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	155	86%
6	6-Jan	155	87%
7	7-Jan	180	89%
8	8-Jan	180	90%
9	9-Jan	180	92%



# Missing Data Solution

## Linear Interpolation

- Interpolation is a mathematical method that adjusts a function to data and uses this function to extrapolate the missing data
- The simplest type of interpolation is the linear interpolation, that makes a mean between the values before the missing data and the value after.
- We could have a pretty complex pattern in data and linear interpolation could not be enough.
- There are several different types of interpolation. Just in Pandas we have the following options like : 'linear', 'time', 'index', 'values', 'nearest', 'zero', 'slinear', 'quadratic', 'cubic', 'polynomial', 'spline', 'piece wise polynomial' and many more .

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	N/A	86%
6	6-Jan	150	87%
7	7-Jan	160	89%
8	8-Jan	N/A	90%
9	9-Jan	180	92%



Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	120	86%
6	6-Jan	150	87%
7	7-Jan	160	89%
8	8-Jan	170	90%
9	9-Jan	180	92%

$$(90+150)/2 = 120$$

$$(160+180)/2 = 170$$

# Missing Data Solution

## Adding a category to capture NA

- This is perhaps the most widely used method of missing data imputation for categorical variables. This method consists in treating missing data as if they were an additional label or category of the variable
- All the missing observations are grouped in the newly created label 'Missing'. It does not assume anything on the missingness of the values. It is very well suited when the number of missing data is high.

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	N/A	99	70%
3	Fast+	167	10%
4	Fast+	90	80%
5	Lite	76	70%
6	N/A	155	10%
7	Fast+	200	95%
8	Lite	76	77%
9	N/A	180	95%



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Missing	99	70%
3	Fast+	167	10%
4	Fast+	90	80%
5	Lite	76	70%
6	Missing	155	10%
7	Fast+	200	95%
8	Lite	76	77%
9	Missing	180	95%

# Missing Data Solution

## Frequent category imputation

- Replacement of missing values by the most frequent category is the equivalent of mean/median imputation. It consists of replacing all occurrences of missing values within a variable by the most frequent label or category of the variable

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	N/A	99	70%
3	Fast+	167	10%
4	Fast+	90	80%
5	Lite	76	70%
6	N/A	155	10%
7	Fast+	200	95%
8	Lite	76	77%
9	N/A	180	95%



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Fast+	99	70%
3	Fast+	167	10%
4	Fast+	90	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	200	95%
8	Lite	76	77%
9	Fast+	180	95%

# Missing Data Solution

## Arbitrary Value Imputation

- Arbitrary value imputation consists of replacing all occurrences of missing values within a variable by an arbitrary value. Ideally arbitrary value should be different from the median/mean/mode, and not within the normal values of the variable.
  - Typically used arbitrary values are 0, 999, -999 (or other combinations of 9's) or -1 (if the distribution is positive).
- 
- This works reasonably well for numerical features that are predominantly positive in value, and for tree-based models in general.
  - This used to be a more common method in the past when the out-of-the box machine learning libraries and algorithms were not very adept at working with missing data.

Dr. Tran Anh Tuan, Faculty of Mathematics and Computer Science, University of Science, HCMC

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%

Arbitrary  
value 999



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	999	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	999	95%
8	Lite	76	77%
9	Fast+	180	95%

# Missing Data Solution

## Adding a variable to capture NA

- When data are not missing completely at random, we can capture the importance of missingness by creating an additional variable indicating whether the data was missing for that observation (1) or not (0).
- The additional variable is a binary variable: it takes only the values 0 and 1, 0 indicating that a value was present for that observation, and 1 indicating that the value was missing for that observation. Typically, mean/median imputation is done together with adding a variable to capture those observations where the data was missing.

Dr. Tran Anh Tuan, Faculty of Mathematics and Computer Science, University of Science, HCMC

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	200	80%
2	Lite	100	70%
3	Fast+	200	10%
4	Fast+	N/A	80%
5	Lite	50	70%
6	Fast+	200	10%
7	Fast+	N/A	95%
8	Lite	200	77%
9	Fast+	180	95%

Median

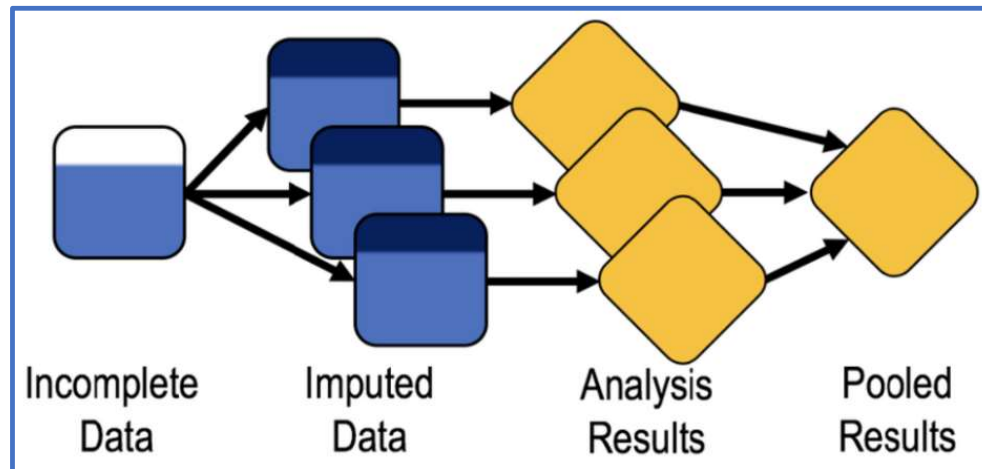
New Feature

Mobile ID	Mobile Package	Download Speed	DL Speed Missing	Data Limit Usage
1	Fast+	200	0	80%
2	Lite	100	0	70%
3	Fast+	200	0	10%
4	Fast+	200	1	80%
5	Lite	50	0	70%
6	Fast+	200	0	10%
7	Fast+	200	1	95%
8	Lite	200	0	77%
9	Fast+	180	0	95%

# Missing Data Solution

## Multiple Imputation

- Multiple Imputation (MI) is a statistical technique for handling missing data. The key concept of MI is to use the distribution of the observed data to estimate a set of plausible values for the missing data.
- Random components are incorporated into these estimated values to show their uncertainty.

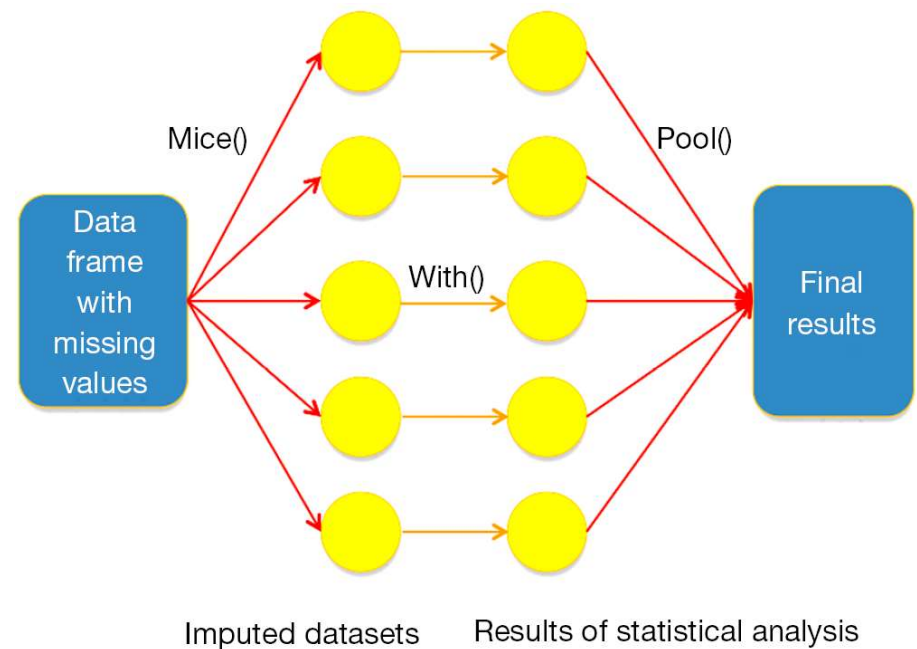




# Missing Data Solution

## Multiple Imputation

- Multiple datasets are created and then analysed individually but identically to obtain a set of parameter estimates. Estimates are combined to obtain a set of parameter estimates.
- As a flexible way of handling more than one missing variable, apply a Multiple Imputation by Chained Equations (MICE) approach. The benefit of the multiple imputation is that in addition to restoring the natural variability of the missing values, it incorporates the uncertainty due to the missing data, which results in a valid statistical inference



## Multiple Imputation by Chained Equations (MICE) – Single Iteration

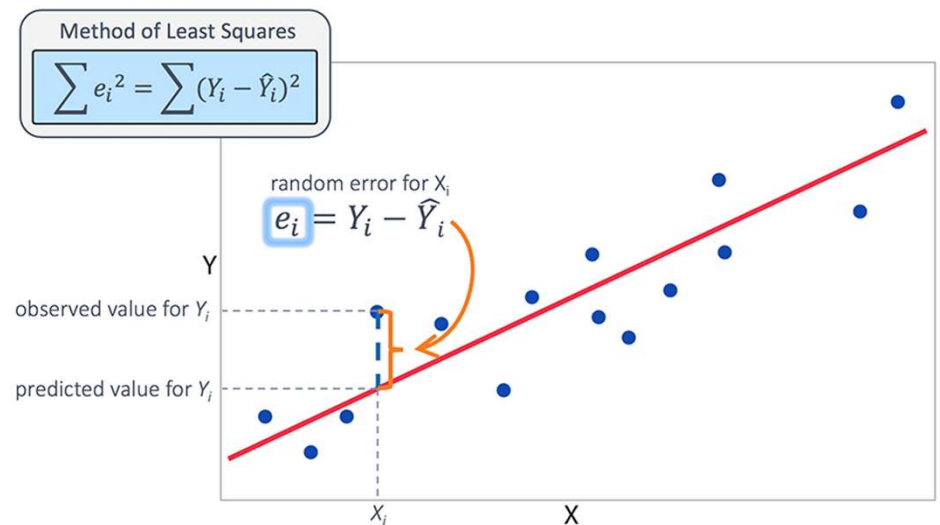


# Missing Data Solution

## Predictive/Statistical models that impute the missing data

- **Linear Regression**

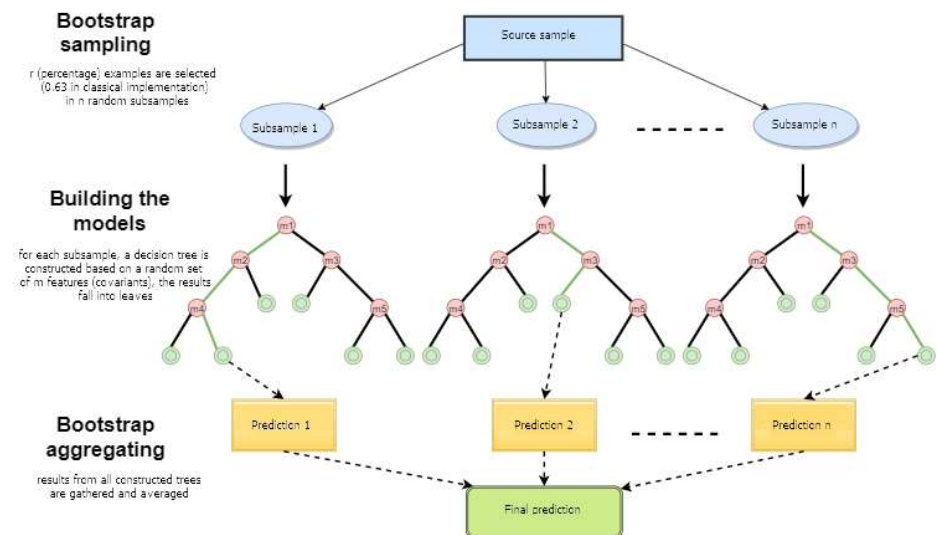
- In regression imputation, the existing variables are used to make a prediction, and then the predicted value is substituted as if an actual obtained value.
- This approach has a number of advantages, because the imputation retains a great deal of data over the list wise or pair wise deletion and avoids significantly altering the standard deviation or the shape of the distribution.



# Missing Data Solution

- **Random Forest**

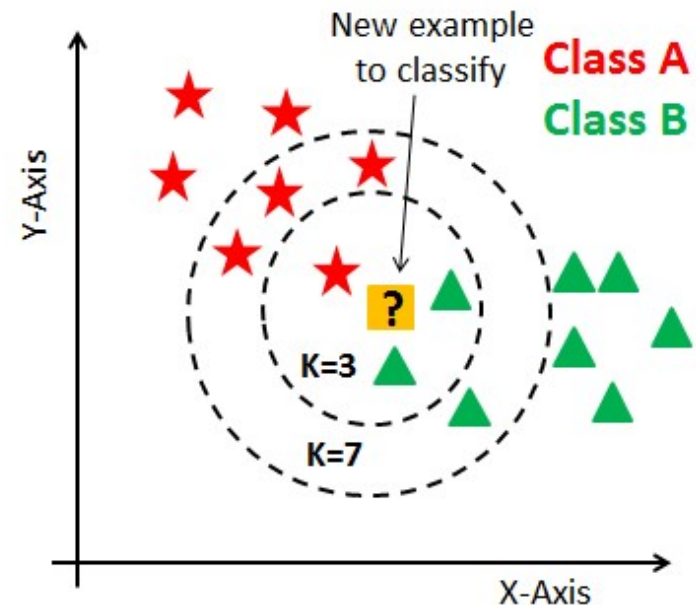
- Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random.
- Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates. One caveat is that random forest works best with large datasets and using random forest on small datasets runs the risk of overfitting.



# Missing Data Solution

- **k-NN (k Nearest Neighbour)**

- k-NN imputes the missing attribute values on the basis of nearest K neighbour. Neighbours are determined on the basis of distance measure.
- Once K neighbours are determined, missing value are imputed by taking mean/median or mode of known attribute values of missing attribute.



# Missing Data Solution

- **Maximum likelihood**

- There are a number of strategies using the maximum likelihood method to handle the missing data. In these, the assumption that the observed data are a sample drawn from a multivariate normal distribution is relatively easy to understand.
- After the parameters are estimated using the available data, the missing data are estimated based on the parameters which have just been estimated.

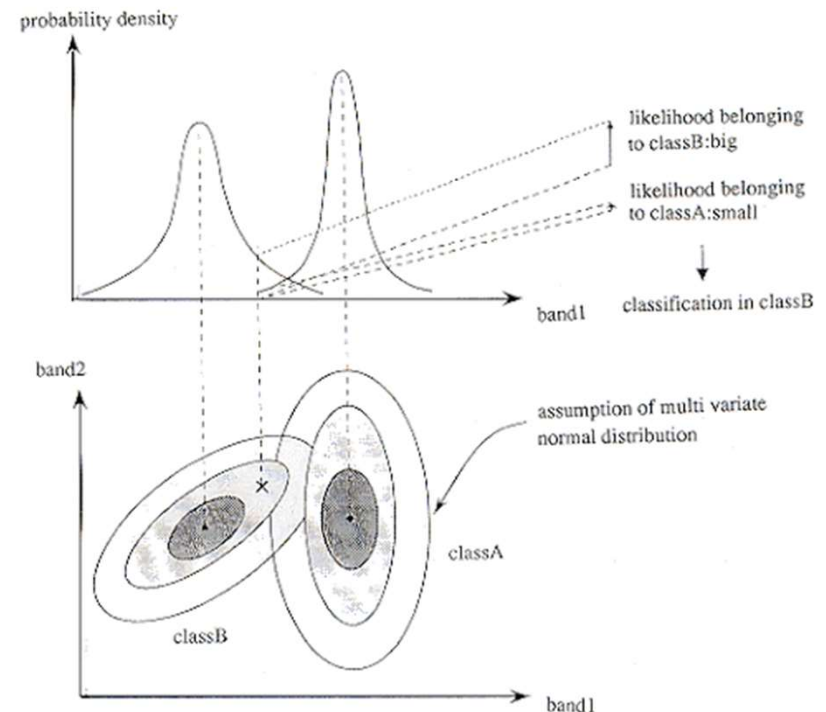


Figure 11.7.1 Concept of Maximum Likelihood Method



# Missing Data Solution

- **Expectation-Maximization**

- Expectation-Maximization (EM) is a type of the maximum likelihood method that can be used to create a new data set, in which all missing values are imputed with values estimated by the maximum likelihood methods.
- This approach begins with the expectation step, during which the parameters (e.g., variances, co-variances, and means) are estimated, perhaps using the list wise deletion. Those estimates are then used to create a regression equation to predict the missing data.

# Missing Data Solution



## • Expectation-Maximization

- The maximization step uses those equations to fill in the missing data. The expectation step is then repeated with the new parameters, where the new regression equations are determined to “fill in” the missing data. The expectation and maximization steps are repeated until the system stabilizes.

Dr. Tran Anh Tuan, Faculty of  
Science, University

### a Maximum likelihood

5 sets, 10 tosses per set

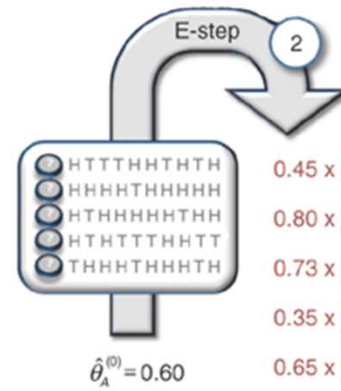
 H T T T H H T H T H  
 H H H H T H H H H H  
 H T H H H H H T H H  
 H T H T T T H H T T  
 T H H H T H H H T H

Coin A	Coin B
	5 H, 5 T
9 H, 1 T	
8 H, 2 T	
	4 H, 6 T
7 H, 3 T	
24 H, 6 T	9 H, 11 T

$$\hat{\theta}_A = \frac{24}{24 + 6} = 0.80$$











$$\hat{\theta}_B = \frac{9}{9 + 11} = 0.45$$

### b Expectation maximization



$$\hat{\theta}_A^{(0)} = 0.60$$

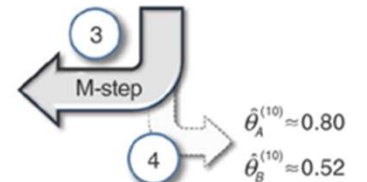
$$\hat{\theta}_B^{(0)} = 0.50$$

0.45 x  0.55 x   
 0.80 x  0.20 x   
 0.73 x  0.27 x   
 0.35 x  0.65 x   
 0.65 x  0.35 x 

Coin A	Coin B
≈ 2.2 H, 2.2 T	≈ 2.8 H, 2.8 T
≈ 7.2 H, 0.8 T	≈ 1.8 H, 0.2 T
≈ 5.9 H, 1.5 T	≈ 2.1 H, 0.5 T
≈ 1.4 H, 2.1 T	≈ 2.6 H, 3.9 T
≈ 4.5 H, 1.9 T	≈ 2.5 H, 1.1 T
≈ 21.3 H, 8.6 T	≈ 11.7 H, 8.4 T

$$\hat{\theta}_A^{(1)} = \frac{21.3}{21.3 + 8.6} \approx 0.71$$

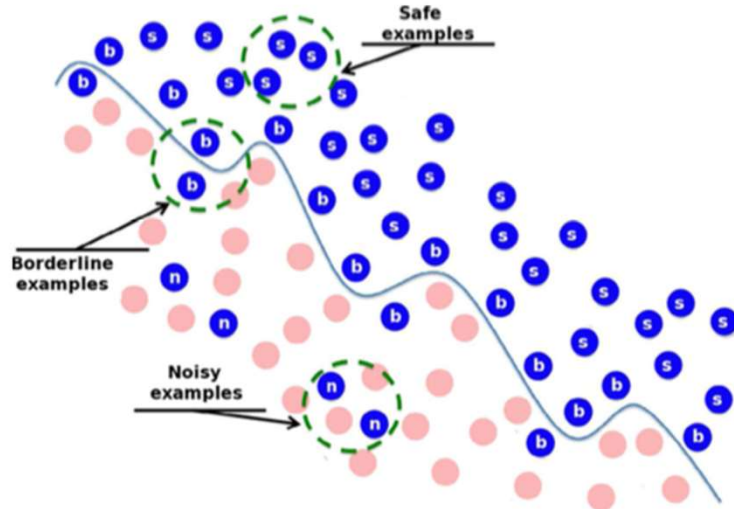
$$\hat{\theta}_B^{(1)} = \frac{11.7}{11.7 + 8.4} \approx 0.58$$



$$\hat{\theta}_A^{(10)} \approx 0.80$$

$$\hat{\theta}_B^{(10)} \approx 0.52$$

# Data Cleaning



See Noisy Data in the next Lecture

## (b). Noisy Data:

Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

### 1. Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

### 2. Regression:

Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

### 3. Clustering:

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.



THANK YOU

Dr. Tran Anh Tuan, Faculty of Mathematics and Computer  
Science, University of Science, HCMC